Linguistic Processing of Text for 
a Large-Scale Conceptual Information Retrieval System

Sung H. Myaeng & Christopher Khoo
School of Information Studies
Syracuse University

Ming Li
Department of Computer & Information Science
Syracuse University

ABSTRACT. This paper describes our large-scale effort to build a conceptual Information Retrieval system that converts a large volume of natural language text into Conceptual Graph representation by means of knowledge-based processing. In order to automatically extract concepts and conceptual relations between concepts from texts, we constructed a knowledge base consisting of over 12,000 case frames for verbs and a large number of other linguistic patterns that reveal conceptual relations. They were used to process a Wall Street Journal database covering a period of three years. We describe our methods for constructing the knowledge base, how the linguistic knowledge is used to process the text, and how the retrieval system makes use of the rich representation of documents and information needs.

1. Introduction

Researchers have attempted to use linguistic processing of various kinds in Information Retrieval (IR) for the past few decades. However, statistical techniques have generally been favored because of their simplicity and efficiency, and because they yield a relatively high performance. There is a general feeling in the IR community that statistical approaches have been fairly thoroughly studied, and little further gain in retrieval performance is likely to accrue from additional effort in developing statistical retrieval techniques.

In the past few years, a number of experimental systems have taken advantage of recent advances in natural language processing and adopted sophisticated linguistic processing with the hope of obtaining a substantial improvement in retrieval performance. Most of these are, however, small-scale experimental systems.

Despite continued uncertainty about whether expensive linguistic processing will pay off for information retrieval, we embarked on a large-scale project (DR-LINK Project) that employs a variety of linguistic processing techniques to enrich the original texts [1], convert both documents and information need statements to a Conceptual Graph representation [2, 3], and retrieve documents at a conceptual level. In this paper, we report our experience in constructing a large linguistic knowledge base for the processing of a large volume of natural language text. We discuss how such knowledge is used to build a rich representation of documents and information need statements, and how the retrieval system utilizes the rich representation.

Our research and development effort, described in this paper, is new in several respects. First of all, the retrieval task for our system is somewhat different from that in conventional retrieval environments where information needs are often simple enough to be expressed in a single sentence or with a set of key words. Our system has to process information need statements consisting of several sentences that explain not only the general areas of interest but also specific constraints that certain concepts must meet. For example, an information need statement may require the location or nationality of the entity involved in an event to be stated in the document. The constraints that are
important for determining relevance cannot be met easily by using keyword-based matching. Sophisticated linguistic processing is needed.

Second, we believe that sophisticated, linguistic processing of text, without a corresponding use of a sophisticated representation scheme for representing concepts and conceptual relations, is not sufficient to improve retrieval performance. Most previous research in the use of linguistic processing techniques has been limited in the sense that the focus has been on improving the quality of index terms by identifying suitable phrases. The same retrieval model (such as the vector model) and the same retrieval techniques as those used with statistically processed text were usually employed in these systems. As a result, regardless of the level of sophistication in text processing, the retrieval performance was bounded by the limitations of whatever term-based model was used. We have strived for a new retrieval technique that exploits a richer text representation.

Compared to the amount of research done in the past for statistically-oriented IR, there has been little accumulation of research on the various ways of applying natural language processing techniques in combination with various retrieval approaches. In particular, we feel that the idea of using conceptual relations for IR has not been explored sufficiently. This project constitutes, among other things, an attempt at using conceptual relations for IR.

Finally, our work can be characterized as one of the recent attempts to carry out IR research with a large volume of data. Much of the experimental work in IR has been limited to relatively small test collections whereas practical systems demand the capability of dealing with substantially larger databases. At least two types of scale-up problems exist: the problem of generalizing the experimental results obtained with small test collections, and the problem of ensuring that the techniques used are practical for handling large amounts of data. The latter problem has been a major criticism against knowledge-based and semantically-based IR systems. Such systems are usually geared toward a limited domain and are, thus, not easily extensible. Knowledge-based methods or linguistic analysis techniques that can be applied to large databases across different domains are necessary.

While different levels of processing (lexical, syntactic, and semantic) are employed for various purposes in our system, this paper focuses on the semantic processing of text, namely, the use of linguistic case frames to convert text automatically into a Conceptual Graph representation. It should be noted that our goal is not just to select high quality indexing words and phrases, but to build Conceptual Graph representation of texts and employ a new retrieval technique developed specifically for this.

2. Conceptual Relations And Information Retrieval

The value of conceptual relations in IR has been studied by various people. Fox [4] and Wang et al.[5] found that retrieval performance could be improved when lexical-semantic relations, that included simple case relations (e.g. AGENT, INSTRUMENT, etc.), were used to find additional terms to expand the original queries.

More direct ways of incorporating relations to improve document and query representations have been investigated from as early as the 1960's. For example, the SYNTOL system used four relations (i.e. CONSECUTIVE, COORDINATIVE, ASSOCIATIVE, and PREDICATIVE) derived directly from automatic syntactic analyses. The direct impact of the approach on retrieval effectiveness was not clear [6].

Relations that were more semantic in nature were used by Farradane [7], who chose nine relations (e.g. ASSOCIATION, FUNCTIONAL DEPENDENCE, etc.) derived from a psychological theory of thinking, in an attempt to improve the representation of